International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Prediction of Air Quality Index of Trivandrum City using Machine Learning Methods

Abishek Jayan

B. Tech Student, Department of Civil Engineering, Mar Baselios College of Engineering and Technology, Trivandrum, India
abishekjayan98[at]gmail.com

Abstract: ***Air Pollution is a global problem that has affected mankind for a very long time. It causes lasting damage to human health and property. As such, governments around the world adopted a system of measuring air pollutant concentrations called Air Quality Index, which provides an easier way to keep track of pollutant concentrations. In this study, we employed two machine learning models, the Extreme Learning Machine model is a variant of the traditional Single Layer Feed forward Artificial Neural Network, which prioritizes speed over accuracy when it comes to making predictions. The Seemingly Unrelated Regression is a traditional statistical model which finds relationships between variables that are uncorrelated with each other but whose error terms correlate, hence the term "seemingly unrelated". The models were trained using three years of data from 2018 - 2020. The optimum combinations of input variables to be used to maximize accuracy were also discovered during this training period. They are then tested for the first three months of 2021. The scoring was evaluated using R^2 scoring method and we observed that the ELM model scored much higher accuracies than the SUR model, making it best suited for predicting the air quality of Trivandrum City.***

Keywords: Air Pollution, Machine Learning, Extreme Learning Machine, Seemingly Unrelated Regression

1. Introduction

We are currently confronted with a slew of environmental issues, including global warming, hazardous waste, resource depletion, air pollution, and a slew of others [1 - 4]. Millions of people die each year as a result of diseases brought on by outdoor air pollution [5]. Air is a necessary component for all living things on the earth. Agricultural burning, volcanic eruptions, and wildfires, as well as urbanization, industrialization, vehicles, power plants, and chemical operations, have all contributed to an increase in pollution over the previous 50 years. Pollution is caused by all of these activities, with particulate matter (PM) being one of the most significant causes [1]. Measured air quality indicators are many times higher above the permissible limit values for human health on a daily basis [5]. According to the Blacksmith Institute's list of the world's most polluted regions from 2008 [1] urban air quality and indoor air pollution are two of the world's biggest pollution challenges. The ever - increasing population, together with its automobiles and companies, is polluting the environment at an alarming rate. The extreme learning machine (ELM) was proposed as a simple learning algorithm for SLFNs based on this concept, which can learn thousands of times faster than traditional feed - forward network learning algorithms such as the back - propagation (BP) algorithm while also achieving better generalization performance. Unlike other learning algorithms, the proposed learning algorithm seeks for the smallest training error and the smallest weight norm. The SUR regression model's main feature is that it uses a set of explanatory variables to define the behavior of a certain research variable. When the objective is to explain the entire system, numerous regression equations may be used. A series of independent linear multiple regression equations, for example, may each represent a different economic phenomenon.

2. Preprocessing

The data was obtained from the website of the Central Pollution Control Board (CPCB). The dataset was then checked for errors and preliminary code setup was done in a Jupyter Notebook. The primary step was to collect pollutant data and meteorological data for training and testing purpose. The pollutant concentrations are dependent on various meteorological factors like ambient temperature, rainfall, Wind speed, relative humidity, solar radiation and bar Pressure. Null values were counted and replaced with the mean of the respective column. As pollutant concentration values were only given in the dataset, extra AQI function were also written in order to convert pollutant concentration values into AQI values based on USEPA guidelines.

3. Extreme Learning Machine

The input weights and hidden layer biases of Single Layer Feed Forward Networks (SLFNs) are randomly selected, and the output weights (connecting the hidden layer to the output layer) of SLFNs may be calculated using a generalized inverse operation of the hidden layer output matrices. The extreme learning machine (ELM) was proposed as a simple learning algorithm for SLFNs based on this concept, which can learn thousands of times faster than traditional feed - forward network learning algorithms such as the back - propagation (BP) algorithm while also achieving better generalisation performance. Unlike other learning algorithms, the proposed learning algorithm seeks for the smallest training error and the smallest weight norm. According to Bartlett's theory of feed - forward neural network generalization performance, the smaller the norm of weights, the better the networks' generalization performance. [3]

Licensed Under Creative Commons Attribution CC BY

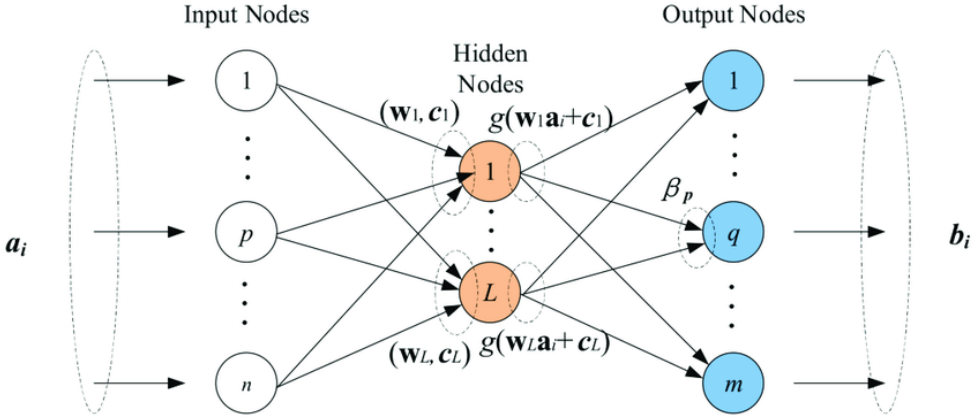

Figure 2.3: ELM Model
(Source: Liu et al., 2018)

4. Seemingly Unrelated Regression

The SUR regression model's main feature is that it uses a set of explanatory variables to define the behaviour of a certain research variable. When the objective is to explain the entire system, numerous regression equations may be used. A series of independent linear multiple regression equations, for example, may each represent a different economic phenomena.

Consider a simultaneous equations model in which one or more of the explanatory variables in one or more equations is also the dependent (endogenous) variable related to another equation in the system. Assume, however, that none of the system's variables are both explanatory and dependent at the same time.

This trend is reflected in the seemingly unrelated regression equations (SURE) model, in which the individual equations are in fact connected to one another, even if they appear to be unrelated at first look.

The jointness of the equations is explained by the structure of the SURE model and the covariance matrix of the associated disturbances. When the individual equations are investigated separately, such joint - ness provides additional information that is not available when the individual equations are studied together. [8]

Assume you have m regression equations.

$$y_{ir} = x_{ir}^{\mathsf{T}} \beta_i + \varepsilon_{ir}, \quad i = 1, \ldots, m.$$

r = 1, 2,. . ., R, where I denotes the equation number.

The number of observations R is considered to be high, but the number of equations m is kept constant.
Each equation I contains a k_i - dimensional vector of regressors x_{ir} and a single response variable y_{ir}.
The model may be represented in vector form by stacking observations corresponding to the i - th equation into R - dimensional vectors and matrices as: -

$$y_i = X_i \beta_i + \varepsilon_i, \quad i = 1, \ldots, m,$$

Where y_i and ε_i are $R \times 1$ vectors, X_i is a $R \times k_i$ matrix, and β_i is a $k_i \times 1$ vector.

5. Methodology

After preprocessing, the models were built and trained with 3 years of data (2018, 2019 and 2020) and was tested against the first three months of 2021. The optimum combinations of inputs to feed into the models were selected for each output. The input combinations for both models vary.

Table 3.1: Optimum Combinations & corresponding accuracies for ELM Model and SUR Model

S. No	Gas/Matter	Optimal Combination of Features (ELM)	R^2 score
1.	$PM_{2.5}$	WD, BP, SR, RH, WS	97.1%
2.	PM_{10}	WD, BP, SR, RH	95.4%
3.	O_3	WD, BP, SR, RH, AT, WS, RF	94.3%
4.	CO	WD, BP, SR, RH, WS	78.3%
5.	NH_3	WD, BP, SR, RH, AT, Temp	75.4%
6.	NO_2	WD, BP, SR, RH, AT, WS	73.3%
7.	SO_2	WD, BP, SR, RH, AT, Temp	65.5%

S. No	Gas/Matter	Optimal Combination of Features (SUR)	R^2 score
1.	$PM_{2.5}$	WD, BP, SR, RH, AT, WS, RF	37.3%
2.	PM_{10}	WD, BP, SR, WS, RF	31.1%
3.	CO	WD, BP, SR	27.1%
4.	NH_3	WD, SR, RH, WS	25.4%
5.	O_3	WD, SR, RH, Temp, WS	24.3%
6.	NO_2	BP, RH, WS	13.3%
7.	SO_2	WD, BP, SR	11.5%

6. Results and Discussions

6.1 General

In the beginning two objectives had been proposed for the project. With regards to the first objective, i. e. building the two models, both the Extreme Learning Machine (ELM) and Seemingly Unrelated Regression (SUR) were built and trained with the same dataset and tested against the same time period. Both of them produced widely varying results as seen from the visualizations below.

Licensed Under Creative Commons Attribution CC BY

6.2 ELM Results

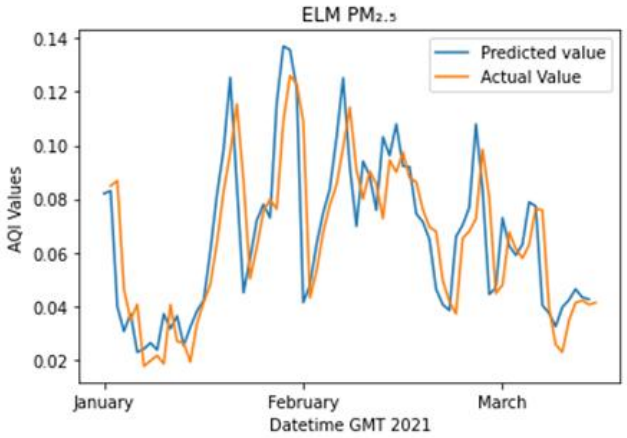

Figure 6.2.1: ELM Comparison of Actual and Predicted AQI for $PM_{2.5}$

In the first test, the prediction of AQI for $PM_{2.5}$ follows closely with the actual values. It is seen that there are only minor variation and disjoints between the predicted values and actual values. The graph is noted to follow the trend and frequency of the graph highly accurately.

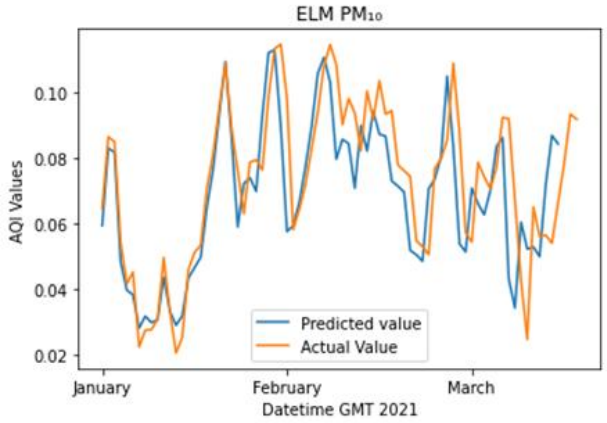

Figure 6.2.2: ELM Comparison of Actual and Predicted AQI for PM_{10}

Similarly, for the second test, the prediction of AQI for PM_{10} also follows suite with the actual values. However, it was noted that during the February - March time period more variation had occurred in the prediction causing it to drift by a small margin from the actual values. The R^2 value proves the accuracy to be extremely high. Overall, this did not affect the performance but further testing is necessary.

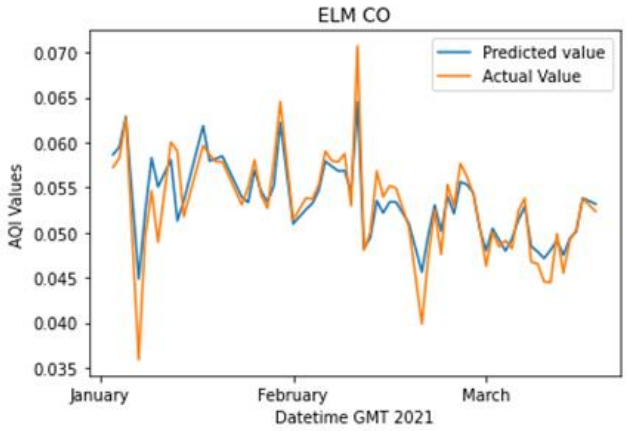

Figure 6.2.3: ELM Comparison of Actual and Predicted AQI for CO

The prediction of AQI for CO is shown to conform to the fluctuations very well throughout the testing period. Minor breakages from the trend were seen occasionally. A clearly visible under prediction was reported during the middle of February. Overall, the visualization proves the high accuracy of the scoring method.

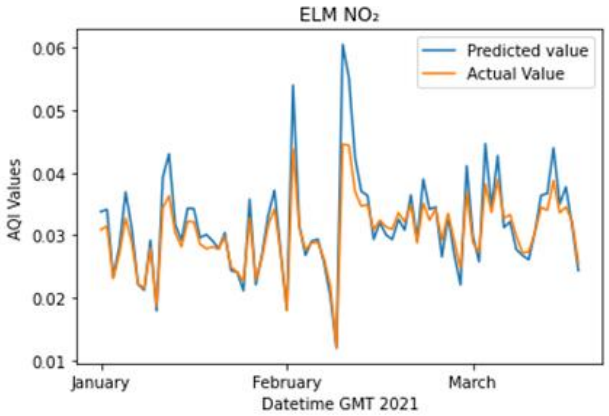

Figure 6.2.4: ELM Comparison of Actual and Predicted AQI for NO_2

Here, the prediction of AQI for NO_2 is reported to follow the actual in the ranges of 0.02 - 0.04. However, the model is seen to over predict when the actual values go beyond that range as clearly seen in the month of February. The overall accuracy is once again seen to conform to the R^2 value.

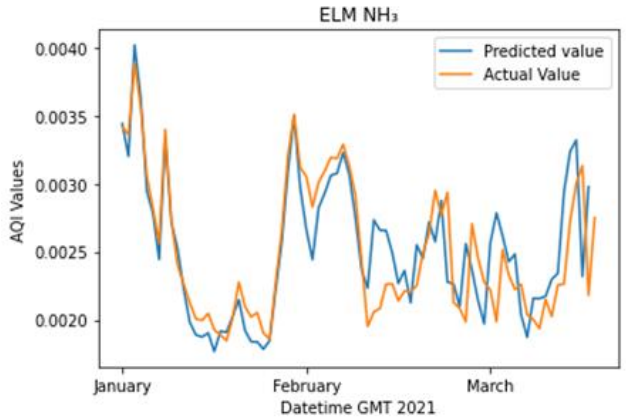

Figure 6.2.5: ELM Comparison of Actual and Predicted AQI for NH_3

Licensed Under Creative Commons Attribution CC BY

For the fifth test, the NH_3 AQI values matches perfectly during January but it is seen that the model tends to under predict when the actual values drop to very low ranges towards the end of January. The variation in the prediction further increases during the February - March time period as seen previously in the PM_{10} visualization.

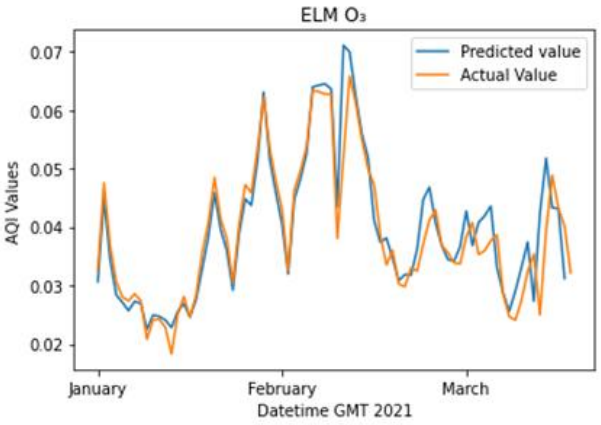

Figure 6.2.6 ELM Comparison of Actual and Predicted AQI for O_3

The O_3 AQI values matches perfectly throughout all three months with a slight variation towards the end of March. Comparing with all the previous visualizations this model combination has shown to be the most accurate in its predictions.

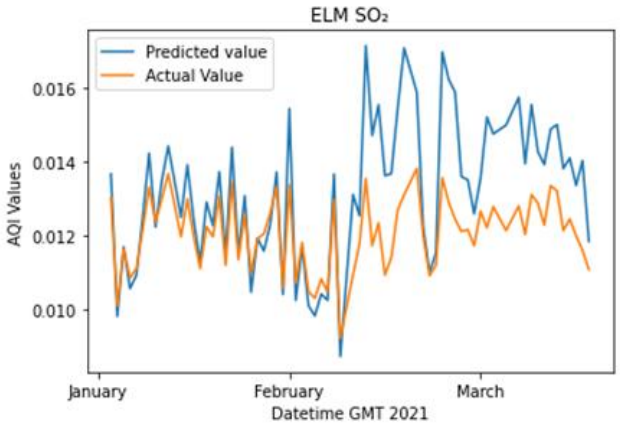

Figure 6.2.7: ELM Comparison of Actual and Predicted AQI for SO_2

Finally, the SO_2 AQI values are reported to have the worst prediction accuracy. The model starts with a very large accuracy, but is seen to have tremendous over prediction towards the latter half of February and beyond. Although, a comeback in matching predictions was seen at the end of February, it is once again observed to fail throughout the month of March.

6.3 SUR Results

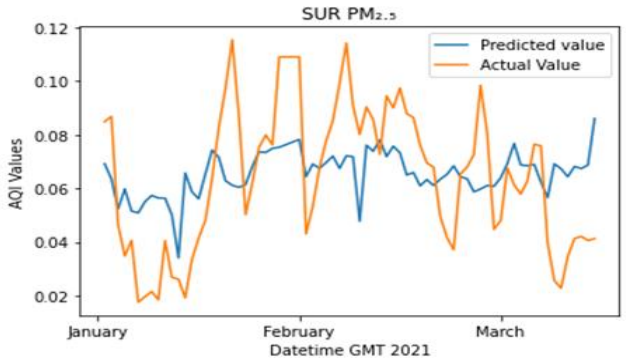

Figure 6.3.1: SUR Comparison of Actual and Predicted AQI for $PM_{2.5}$

Here, the $PM_{2.5}$ predictions are reported to be the best of all of the SUR visualizations. It is seen that there are points throughout the timeline were the prediction and the actual meet and continue to match for a short period. Overall the accuracy is seen to be quite low with a very large gap in prediction towards the end of March.

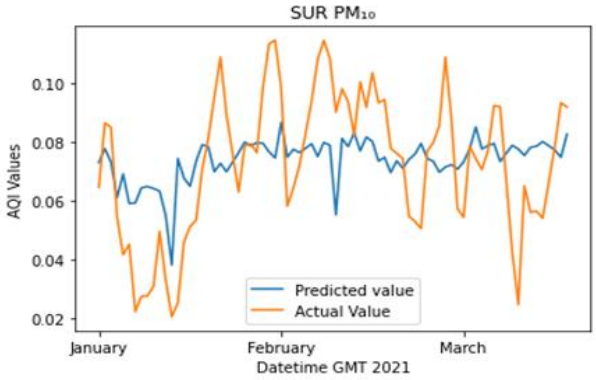

Figure 6.3.2: SUR Comparison of Actual and Predicted AQI for PM_{10}

The AQI prediction for PM_{10} is reported to be lower than $PM_{2.5}$ by a small margin. The data visualization above had shown that the prediction wasn't able to conform to the fluctuations in the AQI completely, although at certain points in seems to follow the variations albeit, for very short periods of time. The predictions are observed to under predict when the actual values go above 0.08 and below 0.04. Overall, it is reported to be similar to the visualization of $PM_{2.5}$ in terms of accuracy.

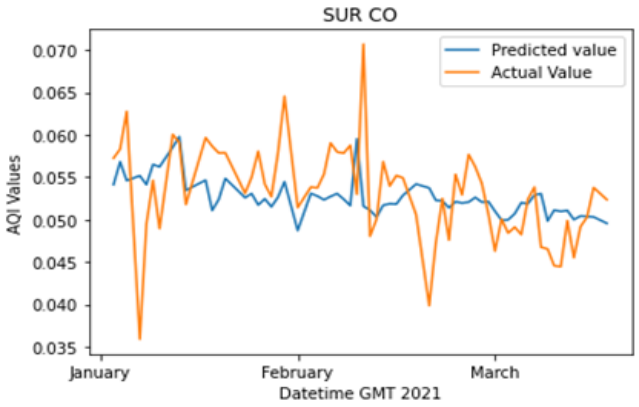

Figure 6.3.3 SUR Comparison of Actual and Predicted AQI for CO

Licensed Under Creative Commons Attribution CC BY

For the third test, the AQI predictions for CO are observed to remain within a range of 0.06 - 0.05 with no outliers. However, it fails to conform to the outliers observed in the early part of January as well as mid - February. The predictions are also noted to miss the trend of variations entirely, as seen in late February.

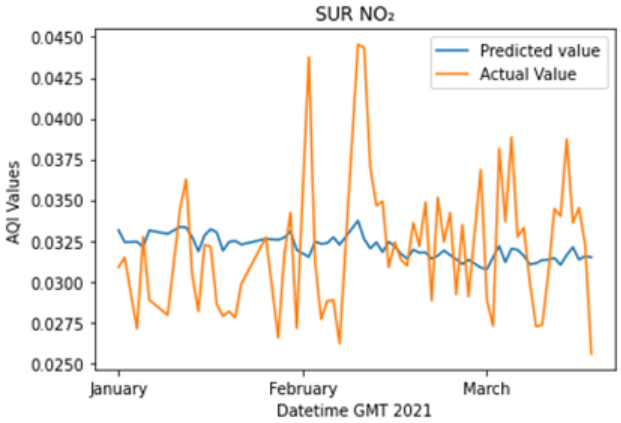

Figure 6.3.4: SUR Comparison of Actual and Predicted AQI for NO_2

Here, the predictions for NO_2 perform very poorly in contrast with the other models. It remains within the range of 0.0325 - 0.0350 and does not fluctuate beyond that range. It is observed to perform very poorly during the latter part of January to mid - February, under predicting the extreme actual data entirely.

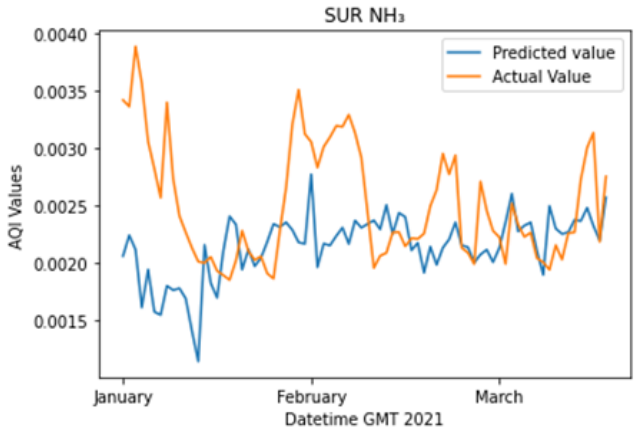

Figure 6.3.5: SUR Comparison of Actual and Predicted AQI for NH_3

It is observed that the predictions for NH_3 begin with wide variations from the actual in early January, but it begins to conform towards the latter part of March. However, the actual and predicted values are seen to conform to the latter part of February through March.

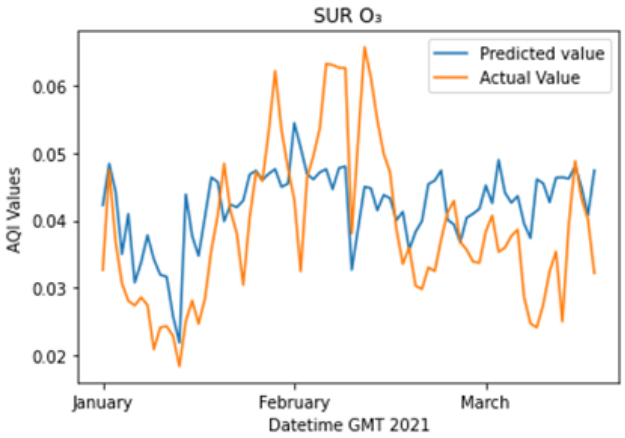

Figure 6.3.6: SUR Comparison of Actual and Predicted AQI for O_3

For the sixth test, the AQI predictions for O_3 are observed to remain within a range of 0.06 - 0.05 with no outliers. However, it fails to conform to the outliers observed in the early part but it begins to conform towards the latter part of March. The initial prediction during the month of January is seen to closely follow the actual data. However, the data is seen to fail that conformity during early February.

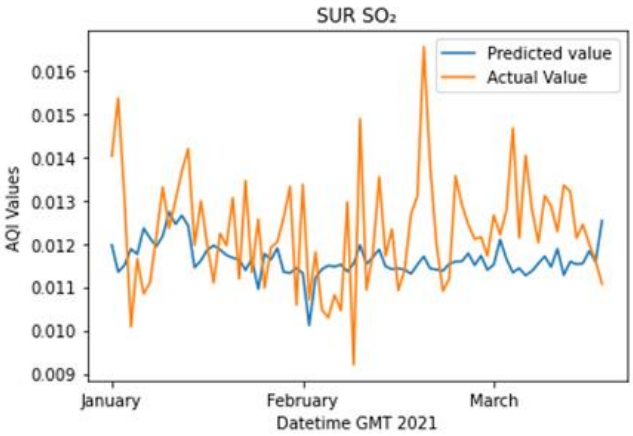

Figure 6.3.7: SUR Comparison of Actual and Predicted AQI for SO_2

Finally, the predicted values of SO_2 are seen to maintain a range from 0.013 - 0.011, as is expected from a statistical model. At one point it is observed to predict below that range but not above. However, the actual data is seen to vary widely from the prediction, despite the general trend being maintained. Overall, the visualized data is seen to fit the calculated R^2 score.

7. Conclusions

As seen from the results, it is concluded that the Extreme Learning Machine (ELM) performs much better than the Seemingly Unrelated Regression (SUR) for all AQI values of particulate matter and gas concentrations.

The Extreme Learning Machine is proven to exceed expectations, scoring above 90% accuracy for the $PM_{2.5}$ and PM_{10} and only drops to a moderate value of 65.5% for SO_2. The predicted AQI values form a very close fit with the

Licensed Under Creative Commons Attribution CC BY

actual AQI values for the months of January, February and March for the year 2021.

However, the Seemingly Unrelated Regression had been reported to have scored poorly, with a maximum accuracy of 37.3% for $PM_{2.5}$ and a minimum of 11.5% for SO_2. But, from the graphs it is clear that while the predicted values do not fit the actual values much in their variation, they do seem to follow the general trend of the data in all cases.

It has been proven from previous machine learning research that the Artificial Neural Network (ANN) model sports high accuracy values for prediction of AQI. For future research, we can consider training and comparing the accuracies for ELM and ANN.

Furthermore, it is possible to improve the future prediction aspect of the project by introducing time series models such as ARIMA, Facebook Prophet Algorithm etc. to predict the meteorological factors for multiple years into the future and using those inputs to get possible predictions for the long - term future.

In conclusion, the Extreme Learning Machine model has proven to be much more capable than the Seemingly Unrelated Regression model at prediction of AQI values and can be considered for further research.

References

[1] Shaharil Mad Saad, Ali Yeon Md Shakaff, A. Saad, A. Yusof, Allan Melvin Andrew, Ammar Zakaria, Abdul Adom. (2017). Development of indoor environmental index: Air quality index and thermal comfort index. AIP Conference Proceedings/ pp.020043

[2] Huixiang Liu, Qing Li, Dongbing Yu, Yu Gu. (2019). Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. Applied Sciences, 9.4069.10.3390/pp.4069

[3] Guang - Bin Huang, Qin - Yu Zhu, Chee - Kheong Siew, Extreme learning machine: Theory and applications, (2006), ISSN 0925 - 2312/pp.489 - 501,

[4] Raghavendra Kumar, Pardeep Kumar, Yugal Kumar. (2020) Time Series Data Prediction using IoT and Machine Learning Technique, Procedia Computer Science, 167. ISSN 1877 - 0509/pp.373 - 381

[5] Kostandina Veljanovska, Angel Dimoski. (2018), Air Quality Index Prediction Using Simple Machine Learning Algorithms, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 7. ISSN 2278 - 6856/pp.025 - 030.

[6] D. L. R., A. C. G., D. Krishnan, R. S. Kumar and S. S. (2020) A Novel Approach for Prediction of Air Pollutant Concentration, *4th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India/pp.217 - 223

[7] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie. (2018) Air Quality Prediction: Big Data and Machine Learning Approaches, International Journal of Environmental Science and Development, 9. /pp.8 - 16

[8] G. Liu, (2015) Seemingly unrelated regression modeling of urban air quality by direct Monte Carlo algorithm, *International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, /pp.171 - 174

[9] Srivastava, Chavi & Singh, Shyamli & Singh, Amit. (2018). Estimation of Air Pollution in Delhi Using Machine Learning Techniques.10.1109/GUCON.2018.8675022/pp.304 - 309

[10] Nayana Vijayaraghavan, Gayathri S Mohan. (2016) Air Pollution Analysis for Kannur City Using Artificial Neural Network, International Journal of Science and Research (IJSR), 5/pp.1399 - 1401

[11] T. M. Amado and J. C. Dela Cruz. (2018) Development of Machine Learning - based Predictive Models for Air Quality Monitoring and Characterization. TENCON 2018 - 2018 IEEE Region 10 Conference, 10.1109/TENCON.2018.8650518/pp.0668 - 0672

Author Profile

Abishek Jayan received his Bachelor's Degree in Civil Engineering from Mar Baselios College of Engineering and Technology, Nalanchira, Trivandrum, Kerala, India. He is currently working as an engineer for a firm.

Licensed Under Creative Commons Attribution CC BY